



# Testing for lack of fit in inverse regression – with applications to photonic imaging

Nicolai Bissantz, Gerda Claeskens, Hajo Holzmann and Axel Munk

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

# Testing for lack of fit in inverse regression – with applications to photonic imaging

Nicolai Bissantz<sup>1</sup>, Gerda Claeskens<sup>2</sup>, Hajo Holzmann<sup>3</sup> and Axel Munk<sup>3</sup>

<sup>1</sup>Lehrstuhl für Stochastik  
Ruhr-Universität Bochum, Germany

<sup>2</sup>ORStat & University Center for Statistics  
K.U.Leuven, Belgium

<sup>3</sup>Institut für Mathematische Stochastik  
Georg-August-University at Göttingen, Germany

March 1, 2007

## Abstract

We propose two test statistics for use in inverse regression problems  $Y = K\theta + \epsilon$ , where  $K$  is a given matrix or operator which cannot be continuously inverted. Thus, only noisy, indirect observations  $Y$  for the function  $\theta$  are available. The tests are designed for hypotheses of the form  $H_0 : \theta(x) = \sum_{j=1}^p a_j \phi_j(x)$ , where  $(\phi_j)_{j \geq 1}$  is the orthonormal system of basis functions given by the spectral decomposition of the operator  $K$ . Both test statistics have a counterpart in classical hypothesis testing, where they are called the order selection test and the data-driven Neyman smooth test. We also introduce two model selection criteria which extend the classical AIC and BIC to inverse regression problems. In a simulation study we show that the inverse order selection and Neyman smooth tests outperform their direct counterparts in many cases. The theory is motivated by data arising in confocal fluorescence microscopy. Here, images are observed with blurring (modeled as deconvolution) and stochastic error at subsequent times. The aim is then to reduce the signal to noise ratio by averaging over the distinct images. In this context it is relevant to decide whether the images are still equal (or have changed by outside influences such as moving of the object table).

**Keywords:** Hypothesis testing, Inverse problems, Model selection, Nanoscale bioimaging, Nonparametric regression, Order selection.

## 1 Introduction

Statistical estimation theory for ill-posed inverse problems, particularly by nonparametric techniques, has been studied intensively within recent years. Examples which fall into this context include deconvolution problems (cf. Fan, 1991, and Johnstone et al., 2004), positron emission and X-ray tomography (Johnstone and Silverman, 1991, and Cavalier, 2000), the heat equation (Mair and Ruymgaart, 1996), imaging (Kaipio and Somersalo, 2005) or problems related to satellite gradiometry (Bissantz et al., 2006).

In these models the object of interest, i.e. the unknown density or regression function  $\theta$ , is

---

<sup>1</sup>Address for correspondence: Dr. Nicolai Bissantz, Faculty of Mathematics, Ruhr-University Bochum, Universitätsstraße 150, Mathematik III NA 3/70, D-44780 Bochum, Germany, email: nicolai.bissantz@rub.de, Phone: +49/234/32-23291, Fax: +49/234/32-14559

not observed directly but only after an application of an operator  $K$ , which for simplicity is assumed to be known in this paper. More specifically, suppose that we have observations  $(z_k, Y_k)$ ,  $k = 1, \dots, n$ , from the model

$$Y_k = (K\theta)(z_k) + \epsilon_k, \quad (1)$$

where the  $z_k$  are fixed design points, the  $\epsilon_k$ 's are i.i.d. errors with  $E\epsilon_k = 0$ ,  $E\epsilon_k^2 = \sigma^2$  and  $E\epsilon_k^4 < \infty$ , and  $K$  is a compact injective operator between  $L_2$ -spaces  $L_2(\mu_1)$  and  $L_2(\mu_2)$ , for measures  $\mu_1, \mu_2$ .

Much work focuses on optimal estimation of  $\theta$  in a mean integrated square error sense, cf. Mair and Ruymgaart (1996), Cavalier and Tsybakov (2002) and Hoffmann and Reiß (2007), among others. Here  $\theta$  is assumed to belong to a certain smoothness class related to the operator  $K$ , or equivalently to satisfy a source condition of type  $\theta \in \text{Im}[\Lambda(K^*K)]$ , where typically either  $\Lambda(t) = t^\nu$  or  $\Lambda(t) = (-\log t)^p$ .

It is well-known that due to the ill-posedness of the problem, one obtains slower rates of convergence for estimating the unknown signal  $\theta$  than for nonparametric estimates with direct observations. The actual rate of convergence is determined both by the smoothness of  $\theta$  as well as by the degree of ill-posedness of the inverse problem. Roughly speaking, for so-called mildly ill-posed problems, one still has a polynomial rate of convergence, while for severely ill-posed problems (including Gaussian deconvolution), in general only logarithmic rates can be obtained. For extensive discussion see Bissantz et al. (2006).

Since nonparametric estimation of  $\theta$  in inverse problems is hard, parametric modeling may become particularly interesting, whenever feasible. However, before actually employing parametric models, one should assess their adequacy via a lack-of-fit (or goodness-of-fit) test. For deconvolution density estimation,  $L_2$ -type statistics for testing goodness-of-fit are discussed in Holzmann et al. (2007). An application to parametric modeling of the brightness distribution in the Milky Way is given in Bissantz et al. (2003). Yet, to our knowledge, for inverse regression models there are no specific methods available. In this paper we shall provide such methods.

We construct omnibus tests, these are tests which are consistent against a wide class of alternative models, by extending the concept of order selection tests and Neyman smooth type tests to the setting of inverse regression models. Lack-of-fit testing based on selecting an appropriate order of an orthogonal series expansion was studied first by Eubank and Hart (1992) and was termed order selection testing, since indeed one form of the test statistic can be viewed as a test on the selected (or estimated) order of the series. Its area of application was extended from linear regression models to likelihood based models by Aerts, Claeskens and Hart (1999, 2000). The selected order in such a test is obtained via a modified version of Akaike's (1973) information criterion. A similar type of test, though originally introduced for testing the distribution function in a goodness-of-fit setting, uses instead the Bayesian information criterion (Schwarz, 1978). These tests build on the idea of a Neyman (1937) smooth test and were introduced by Ledwina (1994). Both the order selection test and the Neyman smooth test extend naturally to inverse regression modeling, where the orthogonal series expansion is canonically given by the singular value expansion, and the ordering of the singular functions is determined by the magnitude of the corresponding singular values.

We study tests for the null hypothesis that the regression function  $\theta$  lies in some finite-

dimensional subspace of  $L^2(\mu_1)$ , i.e. that

$$H_0 : \theta(z) = \sum_{j=1}^p a_j t_j(z), \quad (2)$$

where the  $t_j$  are given functions in  $L^2(\mu_1)$ . Note that since  $K$  is injective, we could equivalently test the hypothesis  $H'_0$  that

$$H'_0 : (K\theta)(z) = \sum_{j=1}^p a_j K(t_j)(z).$$

Since the integral operator  $K$  and the functions  $t_j$  are known, the hypothesis  $H'_0$  is completely specified. Further,  $K\theta$  follows a direct regression model, and therefore the above mentioned methods could be applied to test  $H'_0$ . Nevertheless, there are several reasons why it is preferable to test the hypothesis  $H_0$  directly. First,  $H_0$  and not  $H'_0$  is the natural way to formulate the hypothesis. Second, deviations from the hypothesis  $H_0$  might be somewhat easier to detect, since an extra application of the integral operator  $K$  typically has a smoothing effect. Thus testing  $H_0$  directly may result in a more efficient test (cf. also Holzmann et al., 2007, for a theoretical investigation and simulation results in density deconvolution). Moreover, testing  $H_0$  directly allows testing in situations where  $K$  is unknown (but has to be estimated as well) and where the hypothesis  $H'_0$  is therefore not completely determined.

We consider testing specific linear hypotheses, namely where the regression function is assumed to be a finite linear combination of certain singular functions, i.e. the basis functions of  $L_2(\mu_1)$  occurring in the singular value decomposition of  $K$ . There are several situations where this is of particular interest. One such example is when  $K$  is self-adjoint and its eigenfunctions are trigonometric basis functions. In this case, testing the hypothesis  $H_0$  given in (2) amounts to testing whether  $\theta$  only has finitely many frequencies, where  $p$  determines the maximal frequency allowed in  $\theta$  under  $H_0$ . In case of rejecting  $H_0$ , the modified order selection test will also provide an alternative estimate of the maximal frequency contained in  $\theta$ .

The paper is organized as follows. In Section 2 we introduce the singular value decomposition and formulate the hypotheses that we are going to investigate. Further, we discuss examples for model (1). Section 3.1 contains a version of the order-selection test for inverse regression models, which was originally introduced for direct regression by Eubank and Hart (1992). In Section 3.2 we construct an inverse data-driven Neyman smooth test, as proposed by Ledwina (1994) for direct density testing. In a simulation study in Section 4 we investigate the power properties of our methods, as compared to tests based on the direct hypothesis  $H'_0$ . It turns out that against alternatives which contain additional eigenfunctions (or more precisely, functions from the singular value decomposition), our methods have a significantly higher power than methods based on  $H'_0$ .

Our study is motivated by a problem in fluorescence nanoscale microscopy. Here, images are observed with blurring and stochastic noise, and the aim is to detect structural changes in this sequence of images. Thus, in Section 5 we extend our methodology to test whether the distinct images observed in this deconvolution problem are equal up to possibly different intensities.

Both the order selection test and the data-driven Neyman smooth test are closely related to model selection criteria, specifically to the Akaike information criterion (AIC) and the

Bayesian information criterion (BIC), and in Section 6 we discuss extensions of our procedures to model selection. All proofs are deferred to the appendix.

## 2 Inverse regression models

### 2.1 Singular value decomposition

Since  $K$  in model (1) is assumed to be compact, we can consider its singular value decomposition. There exist orthonormal bases  $(\phi_j)$  of  $L_2(\mu_1)$  and  $(\psi_j)$  of  $L_2(\mu_2)$ , and singular values  $\lambda_j > 0$ , such that  $K\phi_j = \lambda_j\psi_j$  and  $K^*\psi_j = \lambda_j\phi_j$ . Here  $K^*$  denotes the adjoint operator of  $K$ . We shall assume that the inverse problem (1) is mildly ill-posed. More precisely, we suppose that there exist  $c, C, \beta > 0$  such that

$$cj^{-\beta} \leq \lambda_j \leq Cj^{-\beta}, \quad j \geq 1. \quad (3)$$

The hypotheses that we are interested in are of the form

$$H_0 : \theta = \sum_{j=1}^p a_j \phi_j, \quad (4)$$

for some (fixed)  $p$ . Note that by orthogonality of the  $(\phi_j)$ ,  $a_j = \langle \theta, \phi_j \rangle_{L_2(\mu_1)}$ . The hypothesis  $H'_0$  now takes the form

$$H'_0 : K\theta = \sum_{j=1}^p b_j \psi_j,$$

where  $b_j = \langle K\theta, \psi_j \rangle_{L_2(\mu_2)}$  satisfies  $b_j = \lambda_j a_j$ . If it is assumed that the design points  $z_k$  are approximately uniform with respect to the measure  $\mu_2$ , an estimate for  $b_j$  is given by

$$\hat{b}_{j,n} = \frac{1}{n} \sum_{k=1}^n \psi_j(z_k) Y_k.$$

In this paper we impose for simplicity the following

**Assumption 1.** Assume that

$$\frac{1}{n} \sum_{k=1}^n \psi_i(z_k) \psi_j(z_k) = \delta_{i,j}, \quad i, j = 1, \dots, n,$$

with  $\delta_{i,j}$  the Kronecker delta.

Under hypothesis  $H_0$ , see (4), and Assumption 1, for  $j = p+1, \dots, n$ ,

$$\hat{b}_{j,n} = \frac{1}{n} \sum_{k=1}^n \psi_j(z_k) \epsilon_k.$$

Further,

$$E\hat{b}_{j,n} = 0, \quad E(\hat{b}_{i,n} \hat{b}_{j,n}) = \delta_{i,j} \sigma^2 / n, \quad i, j = p+1, \dots, n.$$

Then an estimator for  $a_j$  is given by

$$\hat{a}_{j,n} = \hat{b}_{j,n} / \lambda_j.$$

The tests suggested in Section 3 will be based on the magnitude of the  $\hat{a}_{j,n}$ .

## 2.2 Examples

**Example 1 (Deconvolution).** Suppose that  $\theta, g \in L^2[0, 1]$  are periodic functions, and consider the convolution operator

$$(K\theta)(z) = g * \theta(z) = \int_0^1 g(z - t)\theta(t) dt.$$

First consider the subspace of functions in  $L^2[0, 1]$  which are symmetric around  $1/2$  (in the following denoted by  $L_s^2[0, 1]$ ), and suppose that  $\theta, g \in L_s^2[0, 1]$ . The operator  $K$  is self-adjoint on  $L_s^2[0, 1]$ , has eigenfunctions  $\phi_j(t) = \sqrt{2} \cos(2(j-1)\pi t)$ ,  $j \geq 1$ , and eigenvalues

$$\lambda_j = \int_0^1 g(t)\phi_j(t) dt, \quad j \geq 1.$$

These are the non-zero Fourier coefficients of  $g$ , for which we assume (3), which implies that  $g$  is of finite smoothness. An orthogonal design is e.g. given by the uniform design

$$z_k = (k - 1/2)/n, \quad k = 1, \dots, n. \quad (5)$$

The hypothesis  $H_0$  means that we test for finite frequencies in the signal  $\theta$ .

If one drops the symmetry assumption on  $g$  (and on  $\theta$ ), the example gets more difficult, since it involves complex eigenfunctions. In fact, on the space  $L_{\mathbb{C}}^2[0, 1]$  of complex-valued square-integrable functions, the operator  $K$  is a normal operator for general (real-valued)  $g \in L^2[0, 1]$  with eigenfunctions  $\phi_j(t) = \exp(2\pi i j t)$ ,  $j \in \mathbb{Z}$ , and eigenvalues

$$\lambda_j = \int_0^1 g(t) e^{-2\pi i j t} dt, \quad j \in \mathbb{Z}.$$

Here one estimates the Fourier coefficients of  $K\theta$  by

$$\hat{b}_{j,n} = \frac{1}{n} \sum_{k=1}^n Y_k \overline{\phi_j}(z_k),$$

where  $\bar{w}$  denotes the complex conjugate of  $w \in \mathbb{C}$ . Since the observations  $Y_k$  and the convolution function  $g$  are real,  $\lambda_j = \bar{\lambda}_{-j}$  and  $\hat{b}_{j,n} = \bar{\hat{b}}_{-j,n}$ . Therefore, in the test statistics introduced in Section 3, one should only use the values  $|\hat{b}_{j,n}|^2$  or  $|\hat{b}_{j,n}/\lambda_j|^2$  for  $j > p$ , in which case the theory could be developed in an analogous way.

The example generalizes easily to the two- and multidimensional case as well as to the two-sample case with equal blurring function  $g$  and design points  $(z_k)$ , as considered in the application in Section 5.

**Example 2 (Radon Transform and Tomography).** The Radon transform is of substantial practical importance as it describes e.g. the map of a cross-section through a patient's body onto the detector space in computer or emission tomography. For detailed information on computerized tomography see e.g. Natterer (1986). In a statistical framework, emission tomography was studied by Vardi et al. (1985), Johnstone and Silverman (1990) and Golden-shluger and Spokoiny (2006), among others.

Let us describe the setting in which we are going to work. The operator in the Radon transform  $K_R$  can be represented by a linear, injective integral operator which maps a function

in the “brain space”  $L^2(B, \mu_B)$  of emission densities in the patient’s body to the “detector space”  $L^2(D, \mu_D)$ . Here we focus on the two-dimensional case, where  $B$  is the unit disc, parametrized by polar coordinates  $(r, \vartheta)$ , and the  $D$  is parametrized by the angle  $\varphi \in [0, 2\pi)$  of the detected line through the patient’s body, and its impact parameter  $s \in [-1, 1]$ . The Radon operator  $K_R$  is compact, and we can consider its singular value decomposition in order to form hypotheses of the form (2). In more detail, this decomposition is as follows. Consider the subspace  $L_{inv}^2$  of functions in  $L^2(D, \mu_D)$  for which  $f(s, \varphi) = f(-s, \pi + \varphi)$ , and introduce the normalized measures  $\mu_B$  on  $L^2(B, \mu_B)$  and  $\mu_D$  on  $L_{inv}^2$ , where

$$\begin{aligned} d\mu_B(r, \vartheta) &= \frac{r}{\pi} dr d\vartheta \quad \text{for } 0 \leq r \leq 1, 0 \leq \vartheta < 2\pi, \\ &\text{and} \\ d\mu_D(s, \varphi) &= \frac{1}{\pi^2} (1 - s^2)^{1/2} ds d\varphi \quad \text{for } -1 \leq s \leq 1, 0 \leq \varphi < 2\pi. \end{aligned}$$

Both are probability measures. Then the Radon transform  $K_R \theta \in L_{inv}^2$  of a function  $\theta \in L^2(B, \mu_B)$  is given by

$$K_R \theta(s, \varphi) = \frac{1}{2\sqrt{1-s^2}} \int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} \theta(s \cos(\varphi) - t \sin(\varphi), s \sin(\varphi) + t \cos(\varphi)) dt.$$

The singular value decomposition of  $K_R$  is as follows. In brain space, the orthonormal basis functions are given by the Zernike functions  $\phi_{(p,q)}$

$$\phi_{(p,q)}(r, \vartheta) = \sqrt{q+1} \cdot Z_q^{|p|} e^{ip\vartheta}, \quad q = 0, 1, 2, \dots, p = -q, -q+2, \dots, q,$$

where  $Z_m^k$  is the radial Zernike polynomial of degree  $m$  and order  $k$ . The associated basis functions of detector space are based on Chebyshev polynomials of the second kind  $U_q(\cos(\kappa)) = \sin((q+1)\kappa) / \sin(\kappa)$  and are given by the functions

$$\psi_{(p,q)}(s, \phi) = U_q(s) e^{ip\phi}, \quad q = 0, 1, 2, \dots, p = -q, -q+2, \dots, q.$$

Moreover, the singular values of  $K_R$  are  $\lambda_{(p,q)} = (q+1)^{-1/2}$ . Thus, for every  $(p, q) \in \{0 \leq q; p = -q, -q+2, \dots, q\}$ ,

$$K_R \phi_{(p,q)} = \lambda_{(p,q)} \psi_{(p,q)}.$$

The observational model is (1), where the design points  $z_k \in [0, 1] \times [0, 2\pi)$  are an (at least approximately) orthogonal design for the basis functions  $\psi_{(p,q)}$ . The hypotheses (2) now describes the signal  $\theta$  as a combination of finitely many functions  $\phi_{(p,q)}$ , where the value of  $q$  restricts the radial complexity of the signal, whereas  $p$  bounds its absolute angular frequency from above. In a two-sample problem, by taking differences one can also apply the methodology for testing equality of two images observed indirectly under the Radon transform.

### 3 Model testing

#### 3.1 The order selection test

The order selection test, introduced by Eubank and Hart (1992), can directly be applied to test the hypothesis  $H'_0$ . Let

$$T'_n = \max_{1 \leq m \leq n-p} \frac{1}{m} \sum_{j=1}^m \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2},$$

where  $\hat{\sigma}^2$  is a consistent estimate of  $\sigma^2$ , cf. e.g. Munk et al. (2005). Eubank and Hart (1992) show that if  $E\epsilon_1^4 < \infty$  and

$$\frac{1}{n} \sum_{k=1}^n \psi_i^2(z_k) \psi_j^2(z_k) \leq C \quad (6)$$

for some  $C > 0$  and all  $p+1 \leq i, j \leq n$ , under  $H'_0$  we have that

$$T'_n \xrightarrow{\mathcal{L}} T', \quad n \rightarrow \infty, \quad (7)$$

where

$$T' = \sup_{m \geq 1} \frac{1}{m} \sum_{j=1}^m Z_j^2,$$

$Z_1, Z_2, \dots$  are i.i.d. standard normal random variables, and  $\xrightarrow{\mathcal{L}}$  denotes convergence in distribution. According to (7), the hypothesis  $H'_0$  is rejected with asymptotic level  $\alpha > 0$  if the value of  $T'_n$  exceeds the  $(1 - \alpha)$  quantile  $q_{1-\alpha}$  of  $T'$ . For the quantile  $q_{1-\alpha}$ , explicit formulas are available; alternatively it can be determined by simulation. For further details cf. Eubank and Hart (1992) or Hart (1997). Note that the test statistic  $T'_n$  is based on the magnitude of estimates  $\hat{b}_{j,n}$  of the Fourier coefficients  $b_j = \langle K\theta, \psi_j \rangle_{L_2(\mu_2)}$ ,  $j = p+1, \dots, n$ , and thus measures the distance of  $K\theta$  from the functions specified in  $H'_0$ . However, if one wants to detect deviations from the functions specified in the hypothesis  $H_0$ , a test which is directly based on the distance between  $\theta$  and the function in  $H_0$  seems to be more appropriate. Therefore, we suggest a test statistic which is based on the magnitude of estimates  $\hat{a}_{j,n}^2$  of the Fourier coefficients  $\langle \theta, \phi_j \rangle_{L_2(\mu_1)}$ . More precisely, consider

$$T_n = \max_{1 \leq m \leq n-p} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{n \hat{a}_{j+p,n}^2}{\hat{\sigma}^2} = \max_{1 \leq m \leq n-p} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2 \lambda_{j+p}^2},$$

where  $\Lambda_{m,p} = \Lambda_m = \sum_{j=p+1}^{p+m} 1/\lambda_j^2$ , and

$$T = \sup_{m \geq 1} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2}. \quad (8)$$

The distribution of the supremum of the i.i.d. random walk,  $T'$ , is explicitly available (cf. Spitzer, 1956), but the derivation depends heavily on symmetry arguments for the vectors  $(Z_1^2 - 1, \dots, Z_n^2 - 1)$ . Therefore these arguments do not carry over to the case of  $T$  which is a supremum of weighted random walks. However, the following lemma shows that the distribution of  $T$  can be approximated arbitrarily well by simulations.



**Lemma 1.** *Suppose that the singular values satisfy (3). Then for  $x > 1$  and  $n \in \mathbb{N}$ ,*

$$P\left(\sup_{m \geq n} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2} \geq x\right) = O\left(\frac{1}{(x-1)^2 n}\right). \quad (9)$$

Furthermore,  $P(1 \leq T < \infty) = 1$ .

The next theorem shows that asymptotic critical values for  $T_n$  are given by the critical values of  $T$ .

**Theorem 2.** *Suppose that in model (1), Assumption 1 is satisfied. Further suppose that the singular values of  $K$  satisfy (3) and that the singular functions  $(\psi_j)$  satisfy (6). Then under the hypothesis  $H_0$  we have that for  $x > 1$ ,*

$$P(T_n \leq x) \rightarrow P(T \leq x), \quad n \rightarrow \infty. \quad (10)$$

Let  $c_\alpha$  denote the critical value at level  $\alpha$  for the statistic  $T$ . Then the test which rejects when  $T_n > c_\alpha$  is equivalent to the test which rejects when  $\tilde{j}_n > 0$ , where

$$\tilde{j}_n = \arg \max_{1 \leq m \leq n-p} A_n(m) \quad \text{and} \quad A_n(m) = \sum_{j=1}^m \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2 \lambda_{j+p}^2} - c_\alpha \Lambda_m.$$

Therefore, if the null hypothesis is rejected, the value of  $\tilde{j}_n$  automatically suggests an alternative model which includes further basis functions. Similarly to the proof of Theorem 2, it can be shown that  $\tilde{j}_n$  converges in distribution to

$$\tilde{j} = \arg \max_{m \geq 1} \left( \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2} - c_\alpha \Lambda_m \right).$$

### 3.2 A data-driven Neyman smooth test

In a series of papers, Ledwina (1994) and Kallenberg and Ledwina (1995) introduced data-driven versions of Neyman's goodness-of-fit test for testing the fit of a density function. Such tests can also be transferred to the lack-of-fit problem in regression models, cf. Hart (1997). Specifically, we apply them to test the hypothesis  $H'_0$  in the indirect regression model (1). Define the sequence

$$S'_n(k) = \sum_{j=1}^k \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2}, \quad k = 1, \dots, n-p.$$

The test statistic used is  $S'_n(\tilde{k}')$ , where  $\tilde{k}'$  is based on the BIC criterion (Schwarz, 1978),

$$\begin{aligned} B'_n(k) &= \sum_{j=1}^k \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2} - k \log n, \quad k = 1, \dots, n-p, \\ \tilde{k}' &= \arg \max_{1 \leq k \leq n-p} B'_n(k). \end{aligned}$$

Then if  $E\epsilon_1^4 < \infty$  and if (6) holds, one has that  $P(\tilde{k}' = 1) \rightarrow 1$ , and therefore

$$S'_n(\tilde{k}') \xrightarrow{\mathcal{L}} \chi_1^2.$$

Note that in order to apply this test, the choice  $k = 0$  is not allowed. Indeed, if it were allowed to obtain that  $k = 0$ , by consistency of the BIC as a model selection criterion, under the null hypothesis  $H_0'$  it would consistently select  $\tilde{k}' = 0$ , and hence not be usable as a test (see Ledwina, 1994). The hypothesis is thus rejected with asymptotic level  $\alpha$  if  $S'_n(\tilde{k}')$  exceeds the  $1 - \alpha$  quantile of a chi-squared random variable with one degree of freedom.

For the hypothesis  $H_0$ , we propose a modified version of the data-driven Neyman smooth test which is based on estimates  $\hat{a}_{j,n}^2$  of the Fourier coefficients  $\langle \theta, \phi_j \rangle_{L_2(\mu_1)}$ . To this end define

$$S_n(k) = \sum_{j=1}^k \frac{n\hat{b}_{j+p,n}^2}{\hat{\sigma}^2\lambda_{j+p}^2}, \quad k = 1, \dots, n-p,$$

and consider the selection criterion

$$\begin{aligned} B_n(k) &= \sum_{j=1}^k \frac{n\hat{b}_{j+p,n}^2}{\hat{\sigma}^2\lambda_{j+p}^2} - \Lambda_k \log n, \quad k = 1, \dots, n-p, \\ \tilde{k} &= \operatorname{argmax}_{1 \leq k \leq n-p} B_n(k). \end{aligned}$$

**Theorem 3.** *Suppose that in model (1), the eigenvalues satisfy (3) and that the eigenfunctions satisfy (6). Then under the hypothesis  $H_0$  we have that as  $n$  tends to infinity,  $P(\tilde{k} = 1) \rightarrow 1$ , and consequently that*

$$S_n(\tilde{k}) \xrightarrow{\mathcal{L}} Z^2/\lambda_{p+1}^2,$$

where  $Z$  is standard normal.

The critical value of the data driven Neyman smooth test for inverse regression is determined by a chi-squared critical value with one degree of freedom, and by the eigenvalue  $\lambda_{p+1}$ , which corresponds to the first additional eigenvector used under the alternative hypothesis. This behavior is expected for a test for which the order is determined by the BIC, with penalty factor the log of the sample size. Similar to the behavior of  $S_n(\tilde{k})$ , for goodness-of-fit testing, for local alternative models it is the first coefficient in the alternative model which determines the local power properties.

## 4 Simulation study

In this section we describe the results of a simulation study of the (inverse) order selection and the (inverse) Neyman smooth tests. For this purpose, we generate data from model (1), where  $K$  is convolution on  $[0, 1]$ . The kernel of the convolution integral is such that its eigenvalues decay as  $\lambda_j \sim 1/j$  for the system of eigenfunctions  $\phi_j(x) = \sqrt{2} \cos(2\pi jx)$ .

First we discuss the order selection test. In Fig. 1 we compare the empirical distribution functions of the test statistic  $T_n$  for  $n = 50, 500$ , respectively, with the distribution function of the asymptotic statistic  $T$ . The noise terms  $\varepsilon_k$  have standard deviation 0.1 and are distributed either normally or according to a  $t$ -distribution with 10 degrees of freedom. We performed 1000 simulations under the hypothesis  $H_0 : a_j = 0$  for all  $j$ , i.e. hypothesis (2) with  $p = 0$ . Note from the figure that the empirical distribution of  $T_n$  has already converged very closely to its asymptotic shape, given by the distribution of  $T$ , for rather small sample sizes. For the  $t$ -distributed noise, which is the more difficult case, the distribution of  $T_{500}$  is hardly

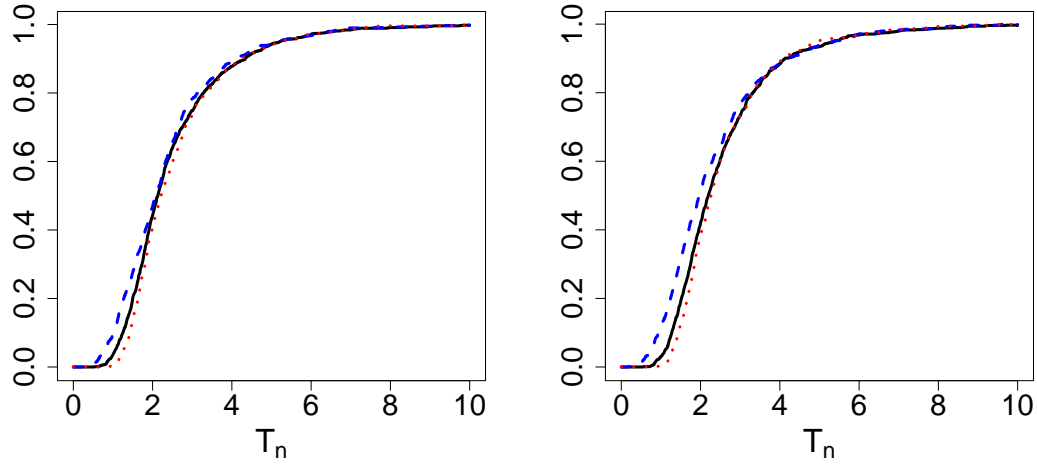


Figure 1: Simulated distribution functions of  $T_{10}$  (dotted line),  $T_{50}$  (dashed line) and  $T$  (solid line) for simulations with normal noise (l.h.s.), and of  $T_{50}$  (dotted line),  $T_{500}$  (dashed line) and  $T$  (solid line) for simulations with t-distributed noise with 10 degrees of freedom (r.h.s.).

Table 1: Simulated power (%) of the classical and inverse order selection tests for levels 0.2, 0.1 and 0.05 from 500 simulation runs. The sample size is either 100 or 500, and  $\sigma$  equals either 0.1 or 0.01. Several choices for the coefficients  $a_5$  and  $a_{10}$  are considered.

$n$	$\sigma$	Level:	Power classical test			Power inverse test		
			0.2	0.1	0.05	0.2	0.1	0.05
100	0.1	$a_5 = 0.010$	99	95	84	100	99	96
		$a_5 = 0.003$	26	11	6	33	20	13
		$a_5 = 0.001$	21	10	6	21	13	6
		$a_{10} = 0.003$	39	15	6	54	36	18
		$a_{10} = 0.001$	25	10	5	26	14	8
500	0.1	$a_5 = 0.001$	28	15	7	32	18	9
		$a_{10} = 0.001$	27	12	7	32	17	11
100	0.01	$a_5 = 0.001$	98	94	82	100	98	98
		$a_5 = 0.0003$	26	13	6	33	19	11

Table 2: Simulated power (%) of the classical and inverse Neyman smooth tests for levels 0.2, 0.1 and 0.05 from 500 simulation runs. The sample size is either 100 or 500, and  $\sigma$  equals either 0.1 or 0.01. Several choices for the coefficients  $a_5$  and  $a_{10}$  are considered.

$n$	$\sigma$	Level:	Power classical test			Power inverse test		
			0.2	0.1	0.05	0.2	0.1	0.05
100	0.1	$a_5 = 0.010$	62	56	55	86	84	83
		$a_5 = 0.003$	24	14	8	28	20	14
		$a_5 = 0.001$	24	15	8	25	16	9
		$a_{10} = 0.003$	24	15	10	41	34	29
		$a_{10} = 0.001$	23	14	8	25	16	10
500	0.1	$a_5 = 0.001$	20	10	6	21	11	7
		$a_{10} = 0.001$	23	14	8	24	15	9
100	0.01	$a_5 = 0.001$	88	87	86	98	98	98
		$a_5 = 0.0003$	24	16	10	29	22	17

distinguishable from the distribution of  $T$ , and for the normal case this is already true for  $T_{50}$ .

In the second part of the simulations we generated data from certain alternatives, where we assume that one of the Fourier coefficients  $a_j$  of  $\theta$  is now nonzero, and the noise is normally distributed. In Table 4 we compare the power of the inverse order selection test with its classical counterpart (where the latter is based on the statistic  $T'_n$ ) for a number of combinations of the design sample size  $n$ , the noise standard deviation  $\sigma$ , and the index  $j$  and value  $a_j$  of a non-zero Fourier coefficient of  $\theta$ . The critical values for the tests for levels of 0.05, 0.1 and 0.2 were determined from the simulations discussed in the preceding paragraph. We conclude that the inverse order selection test outperforms its ordinary counterpart for perturbation amplitudes  $a_j$ , which are below a level where the power is 1 for both the inverse and classical order selection test, and large enough such that the power is still above the significance level of the test.

Finally, we performed simulations of the inverse and classical Neyman smooth tests for the same settings as for the order selection test. The results for the power are given in Table 4, and indicate that the inverse Neyman smooth test outperforms its classical counterpart, in a similar way as for the order selection test.

## 5 Application to confocal fluorescence microscopy

### 5.1 Modelling the observations

Fluorescence microscopy on nanoscales is often hampered by poor signal-to-noise ratios of the individual image frames. Hence, in general a series of several image frames is taken consecutively and accumulated or averaged to yield a single image. However, this has the disadvantage that slight movement of the object table or of the observed object, e.g. in live cell imaging, results in an additional blur of the final image. This could be avoided by testing the individual image frames for significant differences, and, possibly, proper image registra-

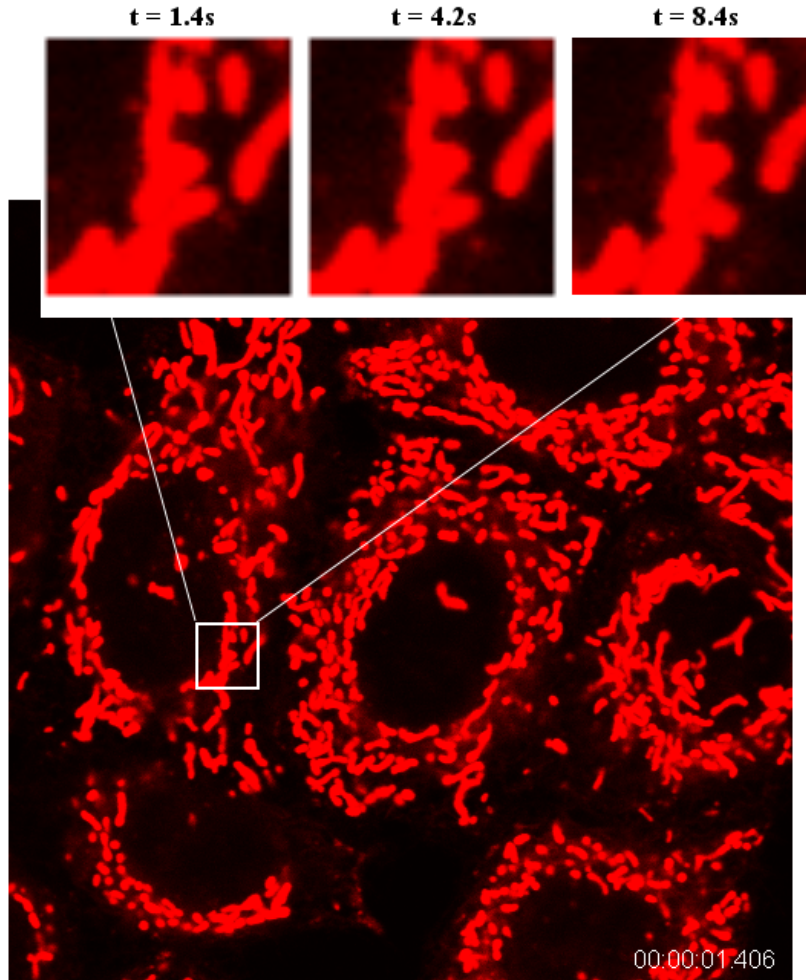


Figure 2: Light microscopic image of fluorescently labeled intracellular membrane structures in HeLa cells (cervix cancer cell line) using the lipophilic carbocyanine dye DiI (red). In the small sub-images we focus on a typical example for alterations in the imaged object due to biological processes in the cells. It shows the movement of a tubular membrane compartment, which may serve as transport or carrier for proteins. Detection of such compartments helps to elucidate intracellular trafficking pathways.

tion before accumulation of the individual frames. Another relevant application is automatic tracking of structural changes within the imaged object over a long time period. For example, GFP (green fluorescent protein)-tagged proteins are used to detect structural changes of intracellular organelles caused by mutations in this protein and for pharmacological tests (cf. Pepperkok and Ellenberg, 2006, for automatic microscopy). Similarly, the monitoring of the motion of intracellular objects such as proteins or transport compartments is of substantial interest. To this end, it is important to note that the image map which describes the microscope is strongly localized, i.e. there is a close correspondence between loci on the observed image and those in the true object. Hence, if we aim to detect changes at certain regions of the cell, it is tempting to restrict the analysis to a suitable part of the image.

In this section we will demonstrate the ability of our inverse testing procedure to detect significant image differences, which can e.g. be used to decide whether single frames may be accumulated straightforwardly. We will first introduce the specific test hypothesis to be used in this application, then introduce the deconvolution model for the data, and finally present the results from an application to a (typical) sequence of images of fluorescently labelled living cells acquired with a confocal laser scanning microscope (Leica TCS).

Our aim is to assess the significance of differences between two images of the same object. Hence, we apply our test to the following null hypothesis:

$$H_0 : I_1 = cI_2$$

where  $I_1$  and  $I_2$  represent the intensity distributions of the first and second image, respectively, and  $c > 0$  is a (real) constant which allows for different exposure times of the images.

Next we introduce the model for such data obtained from confocal microscopy. Typically, available are count data representing observed image intensities on a two- or three-dimensional, equidistant, grid of design-points in the unit square or cube. Here, we consider the two-dimensional case, however the extension to three-dimensional data is straightforward. The design points are given by

$$z_{jk} = \left( \frac{2j-1}{2n}, \frac{2k-1}{2n} \right), \quad 1 \leq j, k \leq n.$$

Hence, at our disposal are observations

$$Y_{jk} = (K\theta)(z_{jk}) + \varepsilon_{jk},$$

where

$$(K\theta)(z) = g * \theta(z) = \int_{[0,1]^2} g(z-t)\theta(t) dt,$$

where “ $*$ ” represents the convolution of the periodic functions  $\theta, g \in L^2([0,1]^2)$ , and  $g$  is called the point-spread-function [PSF] of the microscope. The PSF describes the imaging process and may be computed from the optical properties of the microscope. It represents the image of a point-source observed by the respective microscope. The convolution with the PSF amounts to a smoothing of the original image of the object, where typical smoothing scales are of order  $\approx 100nm$ . However, often the biological structures of interest within the imaged object are of comparable size, and it therefore is necessary to compensate for the convolution of the image with the PSF. Finally, the standard model for the distribution of the photon count data  $Y_{jk}$  is  $Y_{jk} \sim \text{Pois}[(K\theta)(z_{jk})]$ , independently.

We now discuss briefly the properties of the operator  $K$  which is needed to apply the inverse tests. The spectral transform for the convolution operator is given by the Fourier transform on  $[0, 1] \times [0, 1]$ . Hence, the eigenfunctions of  $K$  are  $\phi_{j,k}(t) = \exp(2\pi i((j, k) \cdot t))$ ,  $(j, k \in \mathbb{Z})$  and  $t \in \mathbb{R}^2$ , and its eigenvalues are

$$\lambda_{j,k} = \int_{[0,1]^2} g(t) e^{-2\pi i((j,k) \cdot t)} dt, \quad j, k \in \mathbb{Z}.$$

The Fourier coefficients  $b_{j,k,n}$  of  $K\theta$  can then be estimated by

$$\hat{b}_{j,k,n} = \frac{1}{n} \sum_{l=1}^n Y_l \phi_{j,k}^*(z_l),$$

and those of the original image  $\theta$  by

$$a_{j,k,n} = \frac{b_{j,k,n}}{\lambda_{j,k}}.$$

Thence, the estimator of  $\theta$  reads

$$\hat{\theta}_{M,N,n}(t) = \frac{1}{n} \sum_{j=-M}^M \sum_{k=-N}^N \sum_{l=1}^n \frac{Y_l \exp(-2\pi i((j, k) \cdot z_l))}{\lambda_{j,k}} \exp(2\pi i((j, k) \cdot t)),$$

where  $z_k, t \in [0, 1] \times [0, 1]$ , and  $M, N$  are regularization parameters.

Next we comment on the implementation of the order selection test for the specific setting under consideration here. A similar reasoning holds true for the Neyman smooth test. In contrast to the one-dimensional case considered in Section 3.1 there does not exist a natural ordering of the eigenvalues by the index of the eigenfunctions anymore. We use the following test statistics for the test of the hypothesis  $H_0$  of zero difference between signals, as defined above. It is based on a surrogate to the one-dimensional ordering of the eigenvalues based on the Euclidean norm of their index pairs  $(j, k)$ :

$$T_n = \max_{1 \leq r_l \leq \sqrt{2n^2}} \frac{1}{\Lambda_{r_l}} \sum_{1 \leq j^2 + k^2 \leq r_l^2} \frac{n^2 |\hat{a}_{j,k,n}|^2}{\hat{\sigma}^2},$$

where  $r_l$  is the ordered sequence of values of  $\sqrt{j^2 + k^2}$  for  $-n \leq j, k \leq n$  (with  $j, k \in \mathbb{Z}$ ), and  $\hat{\sigma}^2$  is an estimator of  $\sigma^2 = E\varepsilon^2$ , cf. below. In our computations we used bivariate difference scheme estimator (Munk et al., 2005). The data considered here are real-valued, hence it follows that  $a_{j,k,n} = \bar{a}_{-j,-k,n}$ , and we compare the realized value  $t_n$  of  $T_n$  computed from the data with the quantiles of the simulated distribution of

$$T = \sup_{1 \leq r_l} \frac{1}{\Lambda_{r_l}} \sum_{1 \leq j^2 + k^2 \leq r_l^2} \frac{Z_{j,k}^2}{|\lambda_{j,k}|^2},$$

where  $Z_{j,k} := Z_{-j,-k} \sim N(0, 1)$ , and

$$\Lambda_{r_l} = \sum_{1 \leq j^2 + k^2 \leq r_l^2} \frac{1}{|\lambda_{j,k}|^2}.$$

Time stamp first frame	Time stamp snd frame	$p$ -value
1.4s	2.8s	19%
1.4s	4.2s	11%
1.4s	5.6s	6%
1.4s	7.0s	4%
1.4s	8.4s	3%
7.0s	8.4s	9%

Table 3:  $p$ -values for testing the null hypothesis that two specific images in the time sequence are identical up to a scaling constant and noise.

## 5.2 Application to the HeLa data

We use our method now to compare single image frames in a sequence of confocal images of living HeLa cells, an established cell line. The standard imaging procedure in this case would be to accumulate a certain number, say 4, images of the same object at the same position to improve the signal-to-noise ratio. In our case, there are 9 images which have been acquired in a total time period of  $\approx 11.25$  seconds. Their voxel size is  $138.9nm$  in  $x$  and  $y$ -direction, which corresponds to a zoom level of 2.5. Even if the total image size is  $512 \times 512$  pixels, we focus on a sub-image with size of  $201 \times 201$  pixels which is chosen such that we focus mainly on the HeLa cell at the middle left of fig. 2. Finally, the PSF for the deconvolution has been estimated by fitting a bivariate Gaussian model to an image of a  $200nm$  bead, taking into account the bead shape. Here, the voxel size  $21.3nm$ , which corresponds to zoom level  $16\times$ . Table 3 gives the result of the application of these tests to a number of comparisons of image pairs from this sequence. From the results we find that, for images acquired within  $\approx 3 - 4s$ , the null is not rejected to within a level of 5%, whereas the images are significantly different for image pairs taken at larger time intervals. This implies that care has to be taken before images taken at larger intervals are accumulated. A detailed visual inspection of the image sequence indeed confirms that the images both change w.r.t. large scale movements, and w.r.t. to changes of small scale features such as moving vesicles, which is to be attributed to actual changes of the object. Concluding, the analysis indicates that the test is sensitive enough to detect actually existing image changes, whereas it is not over-sensitive, in the sense that for images which are similar up to noise and exposure time, such as those at timestamps 1.4s and 2.8s the null is not rejected.

The HeLa data has been obtained with a standard confocal laser scanning microscope, where the PSF can be modelled well by a unimodal function. However, higher resolution images (particularly in the  $z$ -direction) can be obtained from more sophisticated fluorescence microscopes such as  $I^5M$  or  $4PI$ -microscopes, where a higher resolution is achieved by a more complicated PSF with strong sidelobes in the  $z$ -direction (cf. Bewersdorf et al., 2006). Therefore, it is here even more important than for confocal microscopy to include the deconvolution into the test statistics. Again, we suggest to use our testing procedure to detect changes in the image due to real changes in living cells, and to find image misalignments prior to image averaging. Related deconvolution settings where image changes and misalignment are important to detect exist in many fields. Here we mention laser ophthalmoscopy (e.g. Nourrit et al., 2005), where it is of interest to detect changes of the human retina (after correction for



image changes due to eye movements).

We close this section with a discussion of some of the assumptions required in Theorems 2 and 3. For our theoretical results we assume the noise to be homoscedastic. In contrast to this, the image data is (approximately) Poisson distributed, and hence the homoscedasticity assumption does not hold. However, the empirical Fourier coefficients  $\hat{b}_{j,k,n}$  are computed from the sum of a very large number of design points (40000 – 250000 depending on the image geometry), which counterbalances this potential problem sufficiently. Indeed, numerical simulations indicate that the test performs well in our data model for a reasonably large number of design points.

Further, the PSF of a fluorescence microscope (e.g. of confocal or *4PI* type) is band-limited, i.e. only has finitely non-zero Fourier coefficients. This implies that it is not possible to attain any information on the object at scales smaller than approximately the Nyquist frequency. Hence, we actually only consider the Fourier transforms of the functions  $\theta$  and  $K\theta$  within the support of the OTF (i.e. the Fourier transform of the PSF), where assumption (3) on the OTF is not necessary anymore.

## 6 Discussion: Connections to model selection and extensions

Both order selection and data-driven Neyman smooth tests have a strong connection to model selection methods. We first define the AIC and BIC in their original framework, then we make a link back to the test statistics, and we proceed by proposing adjusted AIC and BIC versions which can be used in inverse regression problems.

The criteria AIC and BIC are originally derived in a likelihood setting. Let  $\ell_n(\gamma)$  denote the log likelihood function of  $Y_1, \dots, Y_n$ , with unknown parameter vector  $\gamma$ . In the direct normal linear regression setting, for example, where  $Y_i = \beta_1\phi_1(x_i) + \dots + \beta_p\phi_p(x_i) + \varepsilon_i$  and the independent errors  $\varepsilon_i \sim N(0, \sigma^2)$ , the log likelihood function is based on the normal density, and the parameter vector  $\gamma_p = (\sigma^2, \beta_1, \dots, \beta_p)$ . For hypothesis testing, the null hypothesis

$$H_0 : E(Y|x) = \theta(x) = \sum_{j=1}^p \beta_j \phi_j(x)$$

is contrasted with the alternative hypothesis

$$H_a : \theta(x) = \sum_{j=1}^k \beta_j \phi_j(x) \text{ for some } k > p.$$

For each value of  $k = p, p+1, \dots$  there is a corresponding maximized log likelihood function  $\ell_n(\hat{\gamma}_k)$ . Hence, to each of the possible models under  $H_a$ , corresponds a value of the criteria

$$\text{AIC}'(k) = 2\ell_n(\hat{\gamma}_k) - 2(k+1), \quad \text{BIC}'(k) = 2\ell_n(\hat{\gamma}_k) - \log(n)(k+1),$$

where  $k+1 = \text{length}(\gamma_k)$ . Maximizing these criteria is equivalent to maximizing the differences

$$\begin{aligned} \text{aic}'(k) &= \text{AIC}'(k) - \text{AIC}'(p) = 2\{\ell_n(\hat{\gamma}_k) - \ell_n(\hat{\gamma}_p)\} - 2(k-p) \\ \text{bic}'(k) &= \text{BIC}'(k) - \text{BIC}'(p) = 2\{\ell_n(\hat{\gamma}_k) - \ell_n(\hat{\gamma}_p)\} - \log(n)(k-p), \end{aligned}$$

for  $k = p, p+1, \dots$ . We recognize the log likelihood ratio statistic. For normal data, up to a constant not depending on  $k$ , this statistic is equal to  $n\{\log(\text{SSE}_p) - \log(\text{SSE}_k)\}$ , where

the error sum of squares  $\text{SSE}_k = \sum_{i=1}^n (Y_i - \hat{\beta}_1 \phi_1(x_i) - \dots - \hat{\beta}_k \phi_k(x_i))^2$ . The likelihood ratio statistic is first order equivalent to the score statistic  $S_k$ . To define  $S_k$ , denote by  $r_n$  an upper bound for the values of  $k-p$ , and let  $\tilde{\gamma}_k$  be the estimator of  $\gamma_k$  under the null hypothesis, that is, where  $\beta_{p+r} = 0$  for  $r = 1, \dots, r_n$ , hence the vector of length  $p + r_n + 1$ ,  $\tilde{\gamma}_k = (\hat{\gamma}_p, 0, \dots, 0)$ . Further, define  $J(\gamma)$  the corresponding Fisher information matrix, with  $J_{k,k}^{-1}$  its submatrix of dimension  $(k-p) \times (k-p)$  in the lower right corner. Then,

$$S_k = \left( \frac{\partial \ell_n(\tilde{\gamma}_k)}{\partial \beta_{p+r}} \right)_{r=1, \dots, k}^t J_{k,k}^{-1}(\tilde{\gamma}_k) \left( \frac{\partial \ell_n(\tilde{\gamma}_k)}{\partial \beta_{p+r}} \right)_{r=1, \dots, k}.$$

For the regression model as in (1) under the orthogonality assumption (and for now assuming normality)  $S_k = S'_n(k)$  is the score statistic in the direct case. For the data-driven Neyman smooth test the connection to the BIC is now immediate, using the score statistic instead of the log likelihood ratio statistic. For the order selection test, there is a resemblance to the AIC, with a fixed penalty of  $c_\alpha$  times the value  $\Lambda_k$ . In the traditional AIC,  $c_\alpha = 2$ .

For indirect regression problems, the value  $\Lambda_k$  plays the role of the effective number of parameters in the model. This value is arrived at via arguments in the last part of the proof of Lemma 1. Hence, for inverse regression problems, the penalty is not the number of coefficients in the model, but rather the *weighted* number of coefficients, where the weights are the eigenvalues of the integral operator.

Based on the reasoning above, we propose two new model selection criteria that are versions of the AIC and BIC type model selection methods for use in inverse regression problems. The criteria read as follows:

$$\text{aic}(k) = \sum_{j=1}^k \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2 \lambda_{j+p}^2} - 2\Lambda_k, \quad \text{bic}(k) = \sum_{j=1}^k \frac{n \hat{b}_{j+p,n}^2}{\hat{\sigma}^2 \lambda_{j+p}^2} - \log(n)\Lambda_k. \quad (11)$$

The order  $k$  for which the criterion takes on the largest value is selected as the best model order for the given data. To our knowledge only Loubes and Ludeña (2004) deal explicitly with model selection in the context of penalized estimation for (nonlinear) inverse problems. In particular, our criterion  $\text{aic}(k)$  in (11) is related to one of their selection criteria, cf. Remark 3.6 in Loubes and Ludeña (2004). In a related approach, Butucea and Comte (2007) investigate estimation based on model selection of linear functional of  $\theta$  in density deconvolution.

We conclude by mentioning some topics for future work. Even though testing for functions that are linear combinations of eigenfunctions covers a range of interesting applications, including testing for equality of two images, there are situations where other hypothesis tests are called for. In such situations, an orthogonalization approach can be included to make the basis functions that occur in the null hypothesis orthogonal to the eigenfunctions of the kernel, something which imposes both technical and practical complications.

The model selection question above translates to an order selection issue. Other interesting model selection problems arise with multivariate data. For example, when the variable  $z$  is multivariate, interest may lie in selecting components of  $z$  to be included in the model.

## Acknowledgements

Hajo Holzmann and Axel Munk would like to acknowledge financial support of the Deutsche Forschungsgemeinschaft, Grant MU 1230/8-2. The work was initiated during a stay of Gerda

Claeskens in Göttingen, which was financially supported by the DFG Graduiertenkolleg 1023 “Identifikation in Mathematischen Modellen”. The authors are especially grateful to Kathrin Bissantz (University of Münster, Germany) for providing the fluorescence microscopy images, and for help with the biological background.

## References

- Aerts, M., Claeskens, G. and Hart, J. D. (1999), Testing the fit of a parametric function. *J. Am. Statist. Assoc.* **94**, 869–879.
- Aerts, M., Claeskens, G. and Hart, J. D. (2000), Testing lack of fit in multiple regression. *Biometrika* **87**, 405–424.
- Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, B. Petrov and F. Csáki (editors), 267–281, Akadémiai Kiadó, Budapest.
- Bewersdorf, J., Egner, A. and Hell, S. W. (2006), 4Pi Microscopy. In: *Handbook of Biological Confocal Microscopy*, ed. Pawley, J., 561–570, Springer, New York.
- Bhattacharya, R. N. and Rango Rao, R. (1976), *Normal Approximation and Asymptotic Expansions*. Wiley, New York.
- Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.
- Bissantz, N., Munk, A. and Scholz, A. (2003), Parametric versus non-parametric modeling? Statistical evidence based on P-value curves. *Mon. Not. R. Astron. Soc.* **340**, 1190–1198.
- Bissantz, N., Hohage, T., Munk, A. and Ruymgaart, F. (2006), Convergence rates of general regularization methods for statistical inverse problems. *Working Paper, University of Göttingen*.
- Cavalier, L. (2000), Efficient estimation of a density in a problem of tomography. *Ann. Statist.* **28**, 630–647.
- Butucea, C. and Comte, F. (2007), Adaptive estimation of linear functionals in the convolution model and applications. Preprint.
- Cavalier, L. and Tsybakov, A. (2002), Sharp adaptation for inverse problems with random noise, *Probab. Theory Relat. Fields* **123**, 323–354.
- Chow, Y. S. and Teicher, H. (1978), *Probability theory. Independence, interchangeability, martingales*. Springer-Verlag, New York-Heidelberg.
- Eubank, R. L. and Hart, J. D. (1992), Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20**, 1412–1425.
- Fan, J. (1991), On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272.
- Goldenshluger, A. and Spokoiny, V. (2006) Recovering convex edges of an image from noisy tomographic data. *IEEE Trans. Inform. Theory* **52**, 1322–1334.
- Hart, J. D. (1997), *Nonparametric smoothing and lack-of-fit tests*. Springer, New York.
- Hoffmann, M. and Reiß, M. (2007), Nonlinear estimation for linear inverse problems with error in the operator. Preprint.

- Holzmamn, H., Bissantz, N. and Munk, A. (2007), Density testing in a contaminated sample. *J. Multivariate Anal.* **98**, 57–75.
- Johnstone, I. M., Kerkycharian, G., Picard, D. and Raimondo, M. (2004), Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B* **66**, 547–573.
- Johnstone, I. M. and Silverman, B. W. (1990), Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251–280.
- Kaipio, J.-P. and Somersalo, E. (2005), *Computational and Statistical methods for Inverse Problems*. Springer-Verlag, New York.
- Kallenberg, W. C. M. and Ledwina, T. (1995), Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Statist.* **23**, 1594–1608.
- Ledwina, T. (1994), Data-driven version of Neyman’s smooth test of fit. *J. Amer. Statist. Assoc.* **89**, 1000–1005.
- Loubes, J.-M. and Ludeña, C. (2004), Penalized estimators for nonlinear inverse problems. Preprint.
- Mair, B. A. and Ruymgaart, F. H. (1996), Statistical inverse estimation in Hilbert scales, *SIAM J. Appl. Math.* **56**, 1424–1444.
- Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005), On difference based variance estimation in nonparametric regression when the covariate is high dimensional, *J. Roy. Statist. Soc. Ser. B* **67**, 19–41.
- Natterer, F. (1986), *The Mathematics of Computerized Tomography*. B. G. Teubner, Stuttgart.
- Neyman, J. (1937), ‘Smooth’ test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 149–199.
- Nourrit, V., Vohnsen, B. and Artal, P. (2005), Blind deconvolution for high-resolution confocal scanning laser ophthalmoscopy, *J. Opt. A: Pure Appl. Opt.* **7**, 585–592.
- Pepperkok, R. and Ellenberg, J. (2006), High-throughput fluorescence microscopy for systems biology. *Nat. Rev. Mol. Cell. Biol.* **9**, 690–696.
- Petrov, V. V. (1995), *Sequences of Independent Random Variables*, Oxford University Press, Oxford.
- Rice, J. (1984), Bandwidth Choice for Nonparametric Regression, *Ann. Statist.* **12**, 1215–1230.
- Schwarz, G. (1978), Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
- Spitzer, F. (1956), A combinatorial lemma and its applications to probability theory. *Trans. Amer. Math. Soc.* **82**, 323–339.
- Vardi, Y., Shepp, L. A. and Kaufman, L. (1985), A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80**, 8–37.

## Appendix

*Proof of Lemma 1.* Note that from (3), there are  $c_1, C_1 > 0$  such that

$$c_1(m^{2\beta+1} - p^{2\beta+1}) \leq \Lambda_m \leq C_1(m^{2\beta+1} - p^{2\beta+1}).$$

Therefore, from the Hajek-Renyi inequality (cf. Petrov, 1995, p. 53) we have for  $k \geq n$  and  $x > 1$  that

$$\begin{aligned} P\left(\max_{n \leq m \leq k} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2} \geq x\right) &= P\left(\max_{n \leq m \leq k} \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2 - 1}{\lambda_{j+p}^2} \geq x - 1\right) \\ &\leq \frac{1}{(x-1)^2} E(Z_1^2 - 1)^2 \left( \frac{1}{\Lambda_n^2} \sum_{j=1}^n \frac{1}{\lambda_{j+p}^4} + \sum_{j=n+1}^k \frac{1}{\Lambda_j^2 \lambda_{j+p}^4} \right) \\ &\leq \frac{L}{(x-1)^2} \left( \frac{1}{n^{4\beta+2}} \sum_{j=1}^n j^{4\beta} + \sum_{j=n+1}^k \frac{j^{4\beta}}{j^{4\beta+2}} \right) \end{aligned}$$

for some  $L > 0$ . Letting  $k \rightarrow \infty$  proves (9) as well as  $P(T < \infty) = 1$ . The statement  $P(T \geq 1)$  will follow if we show that

$$\frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2} \rightarrow 1 \quad a.s., \quad m \rightarrow \infty.$$

To this end note that the conditions of Theorem 3 (ii), Chap. 3.3. in Chow and Teicher (1978) are satisfied for  $a_j = 1/\lambda_{j+p}^2$ , since  $b_m = \Lambda_m/a_m \sim m$  evidently satisfies the conditions of their theorem. This concludes the proof of the lemma.  $\square$

*Proof of Theorem 2.* Since  $\hat{\sigma}^2$  is consistent for  $\sigma^2$ , it will be sufficient to prove Theorem 2 for the case of known  $\sigma^2$ . We will show that  $P(T_n > x) \rightarrow P(T > x)$  for  $x > 1$ . Let

$$S_{m,n} = \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{n \hat{b}_{j+p,n}^2}{\sigma^2 \lambda_{j+p}^2}, \quad S_m = \frac{1}{\Lambda_m} \sum_{j=1}^m \frac{Z_j^2}{\lambda_{j+p}^2}.$$

Then

$$|P(T_n > x) - P(T > x)| \leq P(A_n) + P(B_n) + |P(C_{n,1}) - P(C_{n,2})|,$$

where

$$\begin{aligned} A_n &= \left\{ S_m \leq x, 1 \leq m \leq m_n, \sup_{m_n+1 \leq m < \infty} S_m > x \right\}, \\ B_n &= \left\{ S_{m,n} \leq x, 1 \leq m \leq m_n, \max_{m_n+1 \leq m < n} S_{m,n} > x \right\}, \\ C_{n,1} &= \left\{ \max_{1 \leq m \leq m_n+1} S_m > x \right\}, \quad C_{n,2} = \left\{ \max_{1 \leq m \leq m_n+1} S_{m,n} > x \right\}, \end{aligned}$$

where  $m_n \rightarrow \infty$  is a sequence which will be specified below. From an application of the multivariate Berry-Esseen theorem (cf. Bhattacharya and Rango Rao, 1976) as in Eubank and Hart (1992) and Hart (1997), we get that

$$|P(C_{n,1}) - P(C_{n,2})| \leq a(m_n) m_n^2 \frac{E\epsilon_1^4}{\sigma^4 \sqrt{n}},$$

where  $a(m)$  only depends on  $m$ . Thus it is possible to find a sequence  $m_n \rightarrow \infty$  slowly enough such that

$$|P(C_{n,1}) - P(C_{n,2})| \rightarrow 0.$$

Furthermore, from Lemma 1 we immediately see that  $P(A_n) \rightarrow 0$ . Finally consider  $B_n$ . Let  $Z_{j,n} = n \hat{b}_{j+p,n}^2 / \sigma^2$ . We follow the proof in Eubank and Hart (1992). We have

$$B_n^c \supset \bigcap_{k=m_n+1}^n \left\{ \frac{|\sum_{j=1}^k (Z_{j,n}^2 - 1) / \lambda_{j+p}^2|}{\Lambda_{k,p}} \leq x - 1 \right\}.$$

Let  $n_j = j^2$ ,  $j(1)$  largest integer  $j$  such that  $j^2 \leq m_n$ , and  $j(2)$  largest integer  $j$  such that  $j^2 < n - 1$ . Define

$$\begin{aligned} \xi_{j,n} &= \max_{1 \leq r \leq n_{j+1} - n_j} \left| \sum_{i=n_j+1}^{n_j+r} (Z_{i,n}^2 - 1) / \lambda_{i+p}^2 \right|, \quad j = j(1), \dots, j(2) - 1, \\ \xi_{j(2),n} &= \max_{1 \leq r \leq n-1-n_{j(2)}} \left| \sum_{i=n_{j(2)}+1}^{n_{j(2)}+r} (Z_{i,n}^2 - 1) / \lambda_{i+p}^2 \right|. \end{aligned}$$

For any integer  $k$  with  $m_n + 1 \leq k \leq n$ , either  $n_j < k \leq n_{j+1}$  for some  $j$  or  $n_{j(2)} < k \leq n - 1$ . It follows that for  $m_n + 1 \leq k \leq n - 1$ ,

$$\frac{|\sum_{i=1}^k (Z_{i,n}^2 - 1) / \lambda_{i+p}^2|}{\Lambda_{k,p}} \leq \frac{|\sum_{i=1}^{n_j} (Z_{i,n}^2 - 1) / \lambda_{i+p}^2|}{\Lambda_{n_j,p}} + \frac{\xi_{j,n}}{\Lambda_{n_j,p}}.$$

Hence

$$B_n^c \supset \bigcap_{j=j(1)}^{j(2)} \left[ \left\{ \frac{|\sum_{i=1}^{n_j} (Z_{i,n}^2 - 1) / \lambda_{i+p}^2|}{\Lambda_{n_j,p}} \leq \frac{x-1}{2} \right\} \cap \left\{ \frac{\xi_{j,n}}{\Lambda_{n_j,p}} \leq \frac{x-1}{2} \right\} \right]. \quad (12)$$

From Markov's inequality,

$$P \left( \bigcup_{j=j(1)}^{j(2)} \left\{ \frac{|\sum_{i=1}^{n_j} (Z_{i,n}^2 - 1) / \lambda_{i+p}^2|}{\Lambda_{n_j,p}} > \frac{x-1}{2} \right\} \right) \leq \frac{4}{(x-1)^2} \sum_{j=j(1)}^{j(2)} \frac{v(1, n_j, n)}{\Lambda_{n_j,p}^2}, \quad (13)$$

where for  $s \leq r$ , by computing the variance of the quadratic form,

$$\begin{aligned} v(s, r, n) &= \text{Var} \left[ \sum_{i=s}^r (Z_{i,n}^2 - 1) / \lambda_{i+p}^2 \right] \\ &= \frac{1}{n^2 \sigma^4} \left( \sum_{k=1}^n \left( \sum_{i=s}^r \psi_{i+p}^2(x_k) / \lambda_{i+p}^2 \right)^2 \right) (E\epsilon_1^4 - 3\sigma^4) \\ &\quad + \frac{2}{n^2} \sum_{k,l=1}^n \left( \sum_{i=s}^r \psi_{i+p}(x_k) \psi_{i+p}(x_l) / \lambda_{i+p}^2 \right)^2 \\ &\leq C \left( \left( \sum_{i=s}^r \frac{1}{\lambda_{i+p}^2} \right)^2 / n + \sum_{i=s}^r \frac{1}{\lambda_{i+p}^4} \right), \end{aligned} \quad (14)$$

where we used (6). Here and in the following  $C > 0$  is a generic constant which might change from line to line. Using (3) one shows that the right-hand side of (13) is of order

$$\sum_{j=j(1)}^{j(2)} \frac{((j^2)^{2\beta+1})^2/n + (j^2)^{4\beta+1}}{(j^2)^{4\beta+2}}, \quad (15)$$

which tends to 0 as  $n, m_n \rightarrow \infty$ .

Next we estimate the contribution of the second term in the intersection (12). Note that from (14), we get

$$v(s, r, n) \leq C \sum_{i=s}^r \frac{1}{\lambda_{i+p}^4}.$$

Therefore we can use an inequality due to Billingsley (1968, p.102) with  $\nu = 2$ ,  $\alpha = 1$  and  $u_i = \lambda_{i+p}^{-4}$  to estimate

$$E\xi_{j,n}^2 \leq C (\log(8j+4))^2 \sum_{i=n_j+1}^{n_{j+1}} \frac{1}{\lambda_{i+p}^4}$$

Thus

$$\begin{aligned} P\left(\bigcap_{j=j(1)}^{j(2)} \left\{\frac{\xi_{j,n}}{\Lambda_{n_j,p}} \leq \frac{x-1}{2}\right\}\right) &\geq 1 - \sum_{j=j(1)}^{j(2)} (\log(8j+4))^2 \Lambda_{n_j,p}^{-2} \sum_{i=n_j+1}^{n_{j+1}} \frac{1}{\lambda_{i+p}^4} \\ &\geq 1 - C \sum_{j=j(1)}^{j(2)} j^{-2}, \end{aligned}$$

which tends to 1 as  $m_n \rightarrow \infty$ . This proves that  $P(B_n) \rightarrow 0$ , as desired.  $\square$

*Proof of Theorem 3.* Using (14) and Markov's inequality, for  $k \geq 2$  we estimate

$$\begin{aligned} P(\tilde{k} = k) &\leq P\left(\sum_{j=1}^k \frac{n\hat{b}_{j+p,n}^2}{\sigma^2 \lambda_{j+p}^2} - \Lambda_k \log n \geq \frac{n\hat{b}_{1+p,n}^2}{\sigma^2 \lambda_{1+p}^2} - \Lambda_1 \log n\right) \\ &= P\left(\sum_{j=2}^k \frac{Z_{j,n}^2 - 1}{\lambda_{j+p}^2} \geq (\Lambda_k - \Lambda_1)(\log n - 1)\right) \\ &\leq C \left( \left( \sum_{i=2}^k \frac{1}{\lambda_{i+p}^2} \right)^2 / n + \sum_{i=2}^k \frac{1}{\lambda_{i+p}^4} \right) / \left( (\Lambda_k - \Lambda_1)(\log n - 1) \right)^2, \end{aligned}$$

which tends to 0 as  $n \rightarrow \infty$ .  $\square$